Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3883–3891
Marseille, 20-25 June 2022

A Generalized Approach to Protest Event Detection in German Local News

Gregor Wiedemann[1], Jan Matti Dollbaum[2], Sebastian Haunss[2], Priska Daphi[3], Larissa Daria Meier[3]
Leibniz Institute for Media Research[1], University of Bremen[2], University of Bielefeld[3]
g.wiedemann@leibniz-hbi.de, dollbaum@uni-bremen.de, haunss@uni-bremen.de,
priska.daphi@uni-bielefeld.de, larissa.meier@uni-bielefeld.de

Abstract

Protest events provide information about social and political conflicts, the state of social cohesion and democratic conflict management, as well as the state of civil society in general. Social scientists are therefore interested in the systematic observation of protest events. With this paper, we release the first German language resource of protest event related article excerpts published in local news outlets. We use this dataset to train and evaluate transformer-based text classifiers to automatically detect relevant newspaper articles. Our best approach reaches a binary F1-score of 93.3 %, which is a promising result for our goal to support political science research. However, in a second experiment, we show that our model does not generalize equally well when applied to data from time periods and localities other than our training sample. To make protest event detection more robust, we test two ways of alternative article preprocessing. First, we find that letting the classifier concentrate on sentences around protest keywords only slightly improves the performance for in-sample data. For out-of-sample data, in contrast, binary F1-scores improve up to +4 percentage points (pp). Second, against our initial intuition, masking of named entities during preprocessing does not improve the generalization of protest event detection models in terms of F1-scores. However, it leads to a significantly improved recall of the models.

Keywords: protest event detection, protest event analysis, text classification, computational social science

1. Introduction

Social scientists conduct protest event analysis (PEA) to learn about developments and trends of the forms, scale and hot topics of political protests to draw conclusions about the state of social cohesion and civil society (Koopmans and Rucht, 2002). For this, large samples of articles from one newspaper are usually selected by a list of key terms, then read manually and judged by relevancy whether they actually contain a protest event. For relevant articles, the protest event information such as the number of participants and topic is extracted. One problem is that keyword-based article selection produces a high number of false positives because indispensable keywords like "demonstrate" (in German: demonstrieren) are ambiguous and can mean either "participate in a protest" or "to show how something works". Weeding out all false positives by reading all articles is very time-consuming. In our project "Protests and Social Cohesion: Comparing Local Conflict Dynamics", we, therefore, strive to conduct PEA in a mostly automatic pipeline of subsequent natural language processing tasks. This paper describes our efforts for the first task, protest event detection.

Our general research goal is to derive a near-complete picture of public protest actions carried out between the year 2000 and the year 2020 in four large German cities (Bremen, Stuttgart, Leipzig, and Dresden). To systematically create variance for comparisons, the cities were selected with several structural criteria in mind, including a population of about the same size, but divergent local political opportunity structures, density of civil society organizations, and their location in the western or eastern part of the country. The project is descriptively interested in the distribution of protest forms, claims, and actors across the given time span, as well as differences between the cities and across the east/west divide. Moreover, from a methodological angle, we plan to compare local protest event data with protest events from national newspapers to better understand possible biases that result from different data sources (Gladun, 2020; Wüest and Lorenzini, 2020; Dollbaum, 2021).

We gather protest events from two different sources. First, where available, we use official data from the local authorities responsible for regulating public affairs. These data, however, have various drawbacks: They are not available for the complete time frame (sometimes due to mandated deletion periods), by their nature they do not contain unregistered protest events, and they often contain only sparse information, lacking, for example, details on organizers and demands. Also, this data necessarily lacks information about the actual progression of protests (number of participants, forms of action, etc.) as it only contains the information, organizers provide before a protest has happened. The second and main source for our protest event data are therefore newspaper articles.

In contrast to most of the literature (Hutter, 2014), we use local rather than national newspapers. This has the advantage to be able to detect many more events and thus gain a much more precise and comprehensive picture (Nam, 2006) that is also less affected by typical news biases such as issue attention cycles (Gladun, 2020). The downside is an increased workload in a procedure that in its traditional form is highly labor-intensive. It was therefore planned from the outset to

automate our annotation procedure. We begin with the first task, to automatically judge the relevancy of newspaper articles based on whether they contain information about a protest event, or not. In this regard, there are two main contributions of our paper:

- We release the first German-language dataset for protest event detection and evaluate several recent transformer-based text classifiers in their ability to perform this task in an automated manner.
- As we expect training data from local news to be highly biased on time and place, we experimentally test two hypotheses regarding alternative preprocessing of the classifier input:
  H1: Focusing the classifier on sentences around keywords improves the model generalizability,
  H2: Replacement of named entities with generic tokens in the training set (masking) improves generalizability.

In the following, we elaborate on related works with a focus on automated approaches to protest event detection (Section 2). Then, we introduce the *German Local Protest News* dataset as the basis for our automated article selection within our project (Section 3). In Section 4, we introduce our experiments with neural network transformers for text classification and test the generalization of our best modeling approach across time and location (Section 5). In Section 6, we give a summary of our findings and an outlook on how we proceed with the results for our overall task of automated protest event analysis.

2. Related works

Human annotation for generating protest event data is labor-intensive and therefore costly (Schrodt and Van Brackle, 2013). Hence, it is difficult to scale, employ in real-time, and transfer to other contexts. Machine coding, by contrast, is not only cheaper but also more transparent and reproducible (Beieler et al., 2016). It is, therefore, no surprise that with declining costs for computational power, automation in the generation of event data has proliferated. Nonetheless, fully automated real-time event coding projects like the International Crisis Early Warning System (ICEWS) and Global Data on Events Language and Tone (GDELT) have shown reliability and validity problems due to large amounts of noise and duplicates (Wang et al., 2016). Higher quality datasets like the Armed Conflict Location & Event Data Project (ACLED) still rely on manual annotation (Raleigh et al., 2010).

Hürriyetoğlu et al. (2019) at CLEF, Hürriyetoğlu et al. (2020) at LREC, and Hürriyetoğlu et al. (2021) at the ACL conference introduced protest event detection as a series of Shared Tasks providing manually annotated articles from South Africa, India, and China to the computer linguistic community. For Germany, a large-scale empirical study was conducted by Wiedemann (2017) who used a combination of rule and dictionary-based text mining to analyze protest event coverage in two national newspapers. As the winning approaches from the Shared Tasks have demonstrated for the English language, much more precise and valid data can be expected by using machine learning, and especially transformer-based models for this task. Currently, many protest event data projects use a combination of machine learning and human coding to reap the benefits of automation while still enjoying the advantages of human annotation where they are most effective.

In generating event data, the part of the workflow that is most often automated is the identification of relevant documents. The decision on relevance has rightly been called a "haystack task" (Hanna, 2017): there are relatively few relevant articles to be found among a large mass of irrelevant ones. This is the case even when documents are pre-selected through a keyword search (Lorenzini et al., 2020; Nam, 2006; Weidmann and Rød, 2019; Wiedemann, 2017); it is all the more urgent in case of random sampling of available newspaper articles for a given period (Hürriyetoğlu et al., 2019).

Given such a strong imbalance of class distribution, standard classification approaches sometimes behave unsatisfactorily when trying to maximize accuracy, as they sometimes falsely label all positive instances as negatives (Croicu and Weidmann, 2015). Solutions thus need to be found that maximize recall, i.e. minimize false negatives. Maximizing recall, however, by definition comes at the price of lower precision (Buckland and Gey, 1994), or the ability to reduce false positives. Prioritizing recall, therefore, means accepting a higher number of false positives that have to be filtered out manually at later stages (Croicu and Weidmann, 2015). Therefore, while the trade-off cannot fully be resolved, one goal is to improve precision to reduce the workload of human coders while retaining as high recall as possible.

A solution has been to work with a second-stage classifier that concentrates on eliminating false positives after an initial round of relevance detection (Zhang and Pan, 2019). Another is to improve the performance of the classifier(s) themselves. For relevance classification, so far, most have used a traditional bag-of-words approach that "considers only patterns of word co-occurrence and proximity and ignores a text's narrative structure" (Nardulli et al., 2015, p. 152). Approaches that take sentence structure into account have been rejected by researchers even a few years ago due to their enormous computational intensity (Nardulli et al., 2015). This situation has changed drastically nowadays with the availability of recurrent and transformer-based neural networks that model not only bag-of-word semantics but also the sequentiality of natural language (Wiedemann and Fedtke, 2021).

Results from automated approaches to relevance detection differ with respect to specific datasets, concrete

models, and reporting styles. Hanna (2017), for instance, reaches F2-scores[1] of up to .77 with an average of .6 across data sources and specifications. Recall is between .47 and .88, precision between .45 and .67. Croicu and Weidmann (2015) use an ensemble classifier that combines different bag-of-words approaches and gets .90 recall and .58 precision for a voting cutoff set at .5. By contrast, the classifier used by Zhang and Pan (2019) for detecting collective action through text and images in social media posts, which is based on a recurrent neural network (RNN) with long short-term memory (LSTM) architecture, achieves .96 and .95 of recall and precision respectively. Likewise, Hürriyetoğlu et al. (2019)'s BERT model produces an F1 score of .9. Recent NLP methods that consider the sentence context thus seem to outperform bag-of-words approaches, which is why we also employ transformer-based neural networks. In general, we conclude from the newest results that protest event detection nowadays performs sufficiently well to substantially speed up protest event analysis as performed in political science.

The latter two works also report results when testing their trained models out-of-sample. In both cases, performance significantly declines: Zhang and Pan (2019) record a decrease to .66 for precision and to .73 for recall. Likewise, when Hürriyetoğlu et al. (2019) test their model on data from a different country than where it was trained (they use English language sources from China and India), the F1-score drops to .64. In a shared task, these numbers could be increased by ten to 15 pp (Hürriyetoğlu et al., 2021), but we nonetheless expect that out-of-sample testing will perform worse than within-sample testing. We deduce from this the necessity to test and potentially improve the generalization of models within our German local news context.

Finally, named entity detection may improve the results. While Croicu and Weidmann (2015) find that it dramatically increases necessary computational power while hardly improving results, Dayanik and Padó (Dayanik and Padó, 2020) on the other hand show that masking named entities actually improves model performance significantly and incidentally also improves out-of-domain performance. We, therefore, consider it an empirical question whether named-entity recognition will work for our task.

3. *German Local Protest News* Dataset

With this paper, we publish the *German Local Protest News* (GLPN) dataset as the first German-language resource of protest-related article excerpts (Wiedemann et al., 2022)[2] published in local news outlets.[3] For each city, the largest local newspaper was selected and accessed through the commercial databases of Factiva, LexisNexis, or Genios (depending on availability). The respective article base was searched with an inclusive search string that reflected the project's broad definition of a protest event, including not only standard terms like "demonstration" or "strike", but also actions like protest performances, occupations, or citizens' initiatives[4]. The resulting numbers of potentially relevant articles vary from 2000 to 10000 per city and year.

To increase efficiency in the manual workflow, we opted for a two-step annotation process. In the *first* step, research assistants were presented with snippets from the extracted articles with keyword in context (KWIC) lists, where the original keywords from the search string were presented with 30 words of article text before and after. Based on this, the research assistants decided whether the article contained the description of a protest event conforming to our definition. Selected articles were then in a *second* step imported into

Year/City	**Leipzig**	**Dresden**	**Stuttgart**	**Bremen**
2009	0	0	771	0
2010	0	0	527	0
2011	0	0	260	0
2012	0	0	209	0
2013	0	0	51	0
2014	0	485	0	0
2015	729	0	0	174
>=2016	391	0	0	361

Datasets	**Not relevant**	**Relevant**	**Total**
training	797	1117	1914
validation	122	152	274
test-*within*	217	330	547
test-*time*	217	535	752
test-*loc*	395	90	485
Total	1748	2224	3972

Table 1: Dataset statistics (background colors indicate how training, validation and three different test sets were compiled from combinations of local newspapers and publication years)

[1]By using F2-scores, the author considers recall twice as important as precision.

[2]Due to copyright issues, we refrain from publishing article full texts. In our experiments, however, we show that full texts are not necessary for protest event detection in newspaper articles and concentration on small, potentially relevant excerpts even improves the automatic approach.

[3]The dataset can be downloaded at: `https://doi.org/10.5281/zenodo.6490537`

[4]The exact search string was: *protest* OR Versammlung* OR demonstr* OR Kundgebung* OR Kampagne* OR "Soziale Bewegung*" OR Hausbesetzung* OR Streik* OR Unterschriftensammlung* OR Hasskriminalität* OR Unruhen* OR Aufruhr* OR Aufstand* OR Boykott* OR Riot* OR Aktivis* OR Widerstand* OR Mobilisierung* OR Bürgerinitiative* OR Bürgerbegehren*.

INCEpTION, a server-based linguistic annotation tool developed at the University of Darmstadt (Klie et al., 2018). In this tool, articles were fully manually annotated with the following variables: date, action form, number of participants, claims, participants and organizers, occurrence of violence, and whether the event is, or triggers, counter-protest.

To create labeled data for supervised machine learning to automate the relevancy decision, we relied on both steps of the manual workflow described above. First, since the manual filtering in step one was done rather inclusively to minimize false negatives it produces a considerable number of false positives. From this subset, we used only those articles as true positives that were confirmed as relevant during the second step of full article annotation. As true negatives, we used those articles that were marked as negatives in step 1, and step 2. During the process of refining the models, it turned out that negatives from step 1 had to be subjected to additional manual corrections, as the model correctly identified protest events that had been overlooked during the initial filtering. Thus, we went through two further iterations of revising the manually assigned codes before arriving at a final training data set. For a representative sample of this set, two coders achieved an agreement of 96.2 % and a Krippendorff's α score of 0.73 for their label decisions.

For the creation of the "German Local Protest News" dataset, we selected from this data collection a subset of 3972 articles. With 56 % relevant articles, this selection contains a considerable bias in favor of relevant articles. We opted for this deviation from the assumed true distribution of classes to reduce the likelihood of false negatives, i.e., overlooked protest events, compared to false positives in a final classification model. Table 1 displays how the sample is distributed across eleven years (2009–2019) and the four German cities (the "Leipziger Volkszeitung" from Leipzig, the "Sächsische Zeitung" from Dresden, the "Stuttgarter Zeitung" from Stuttgart, and the "Weser Kurier" from Bremen). To be able to test the generalizability of our protest event classifiers, we decided to create one training, one validation, and three different test sets:

- *training and validation:* All articles from Leipzig (2015), Stuttgart (2009–2013), and Bremen (2015) are selected and randomly split into 70 % for training, and 10 % for validation.
- *test-within:* The remaining 20 % from the previous selection serve as in-sample test data for our experiments.
- *test-time:* Articles from Leipzig (2016) and Bremen (2017) were selected to create a test set from the same cities as the training data, but with coverage of distinct subsequent time periods to be able to determine the extent to which any change in protest activity over time will affect the performance of automatic event detection.

Parameter	Value
maximum sequence length	512
batch size	4
lr scheduler	linear
warmup ratio	0.1
learning rate	5e-6
weight decay	0.2
number of epochs	6

Table 2: Hyperparameters for transformer fine-tuning

- *test-loc:* Articles from Dresden (2014) were selected to observe how much the performance of automatic event detection is affected if not the time but the place of the target sample changes.

4. Automatic Relevancy Classification

We perform the first experiment on the GLPN dataset to select the best performing machine classifier to perform the identification of relevant documents in an automated manner. For this, we rely on fine-tuning of current pretrained transformer-based language models for the German language, and one multilingual model. Since Devlin et al. (2019) published the BERT model, fine-tuning of pretrained language models, based on variants of the transformer architecture, sets the state of the art for text classification.

For the German language, we identified six different pretrained models based on three transformer variants.

BERT: The standard BERT model is a neural-network transformer pretrained on masked language modeling (MLM) and next sentence prediction as a self-supervised target task on very large unlabeled datasets. For MLM, the model is trained to guess a small number of masked tokens (15 %) in a given training text. By this, it creates contextualized word vectors as internal representations that contain complex semantics that can be used in any downstream NLP task. In 2019, deepset.ai released the first model based on the initial BERT architecture that was trained exclusively on case-preserving German language data (`bert-base`).[5] One year later, they released with `gbert-base` and `gbert-large` updated versions of German BERT models (Chan et al., 2020) that were trained on more and cleaner data. The base and large versions differ in model size, which corresponds to their capacity to learn linguistic and semantic knowledge (base: 110m, large: 335m trainable parameters).

ELECTRA: The ELECTRA model by Clark et al. (2020) modifies the BERT pretraining procedure by replacing the MLM task with a 'replaced token detection' task. For this, instead of masking the input text partially, some words in an input text are replaced with plausible, synthetically generated tokens. The classifier during pretraining then has to discriminate whether or

[5] `https://deepset.ai/german-bert`

Model	Not relevant			Relevant			Macro-avg		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
bert-base	87.9±1.0	88.6±1.4	88.2±0.6	92.5±0.8	92.0±0.8	92.2±0.3	90.2±0.5	90.3±0.5	90.2±0.4
gbert-base	**90.8±1.2**	85.8±1.8	88.2±0.9	91.0±1.0	**94.2±0.9**	92.6±0.5	90.9±0.7	90.0±0.8	90.4±0.7
gbert-large	88.7±1.8	89.2±2.7	88.9±0.7	92.9±1.6	92.5±1.5	92.7±0.4	90.8±0.5	90.9±0.7	90.8±0.5
gelectra-base	85.9±2.9	87.4±3.5	86.5±0.5	91.7±2.1	90.4±2.8	91.0±0.6	88.8±0.7	88.9±0.5	88.8±0.4
gelectra-large	89.8±0.4	90.0±1.3	**89.9±0.7**	**93.4±0.8**	93.2±0.3	**93.3±0.4**	**91.6±0.5**	**91.6±0.7**	**91.6±0.6**
xlm-roberta	86.3±1.3	**90.2±0.6**	88.2±0.5	**93.4±0.3**	90.6±1.1	92.0±0.5	89.9±0.6	90.4±0.4	90.1±0.5

Table 3: Protest event relevance classification performance of transformer models pre-trained on German language data (in %, mean and standard deviation for five repeated runs).

not each token was replaced by the generator process. By this, the language model is able to learn from the entire input sequence instead of from the fraction of masked tokens only. This is supposed to improve the resulting language model in terms of speed and quality. Chan et al. (2020) also released two German ELECTRA models of base and large size.

XLM-RoBERTa: This transformer is trained with the standard MLM objective, but for a hundred languages simultaneously (Conneau et al., 2020). The very large multilingual training corpus with a size of two terabytes was filtered from the CommonCrawl [6] dataset. To deal with all the languages in one model, it scales up the capacity of BERT (550m trainable parameters) and increases the vocabulary size to 250k.

All models are fine-tuned on the GLPN training set with the same reasonable default hyper-parameters (cf. Table 2).[7] The best performing model on the validation set is used for evaluation on the *within-test* set. Table 3 reports the mean test set performance and standard deviation of each model for five repeated runs. The achieved results with a macro-F1 score ranging from 88.8 % for the gelectra-base model to 91.6 % for the gelectra-large model are surprisingly close. However, the ranking of the models is as expected from the literature. The large models perform slightly better than their base counterparts do, and the ELECTRA model is superior to standard BERT. Despite its highest capacity, the performance of the multi-lingual XML-RoBERTa model does not exceed that of the monolingual transformers. For all models, evaluation metrics on the *relevant* category are significantly higher than for the *irrelevant* articles, which is beneficial to our data selection process that prefers reducing false negatives over reducing false positives. We consider the gelectra-large performance of 93.3 % binary F1-score with a slight potential for improvement through further hyperparameter tuning to be very well suited for an application in our automatic protest event detection workflow.

[6] https://commoncrawl.org

[7] Hyperparameter optimization (e.g. on the learning rate, warmup ratio, and weight decay) probably would lead to slightly improved results.

5. Improving Generalizability

Although the previous in-sample evaluation suggested sufficient performance of our approach, we expect performance drops when applying the classifier to out-of-sample data. This is because that linguistic patterns describing local protest events very likely change significantly across time and place because central topics, actors, and even forms of protest change. Thus, there is a need to further evaluate the severity of performance loss of our automation approach regarding these data dimensions. Further, we test two hypotheses on alternative preprocessing of the news articles to improve the generalizability of our model.

5.1. Sentences around Keywords

We observe that protest events often are a side story in articles rather than the main point of news. In these cases, a classifier might get distracted by the content unrelated to the protest event. Even worse, the sequence length of 512 tokens leads to a cut-off of content for longer articles. In this case, a classifier would have no chance to identify a relevant article if the protest event is mentioned at the end of the news story. Due to this circumstance, we tried to extract the potentially most relevant content to the classifier by extracting relevant sentences from an article. For this, each article is separated into sentences first. Then, if sentences match our initial keyword search (cf. Section 3), we keep them. Otherwise, they are removed from the article. Independent of the keyword matching, we keep the headline as the first sentence in the article. To provide the classifier with more, potentially useful context around matching key terms, we further test variants in which we keep an additional +1 or +2 sentences before and after the sentence containing a keyword match.
We fine-tune the best performing transformer model *gelectra-large* analog to the previous experiment but with the altered preprocessing of the input texts. Table 4 displays the results for all preprocessing variants on all three test datasets. Since we are mainly interested in the performance of our approach to identify articles with actual protest event mentions, we report binary evaluation scores with *relevant* as the positive class instead of the macro-F1 scores for the following evaluations.

Model	Precision	Recall	F1-score
		test-within	
fulltext	**93.4±0.8**	93.2±0.3	93.3±0.4
kwic+0	91.7±1.7	94.4±1.1	93.0±1.0
kwic+1	92.6±1.0	**94.9±1.4**	**93.7±0.2**
kwic+2	92.6±1.1	94.7±1.2	93.6±0.3
		test-time	
fulltext	**93.4±0.4**	83.9±1.5	88.4±0.8
kwic+0	90.7±1.7	85.2±5.2	87.7±2.2
kwic+1	92.0±1.6	**85.9±2.6**	88.8±0.9
kwic+2	92.4±0.8	85.6±0.5	**88.9±0.3**
		test-loc	
fulltext	76.6±6.9	70.0±17.2	71.6±9.5
kwic+0	75.0±11.1	62.7±19.5	65.5±6.4
kwic+1	79.2±7.8	**73.3±9.7**	**75.4±2.4**
kwic+2	**79.7±10.3**	69.1±18.0	71.8±6.8

Table 4: Classification performance with full-text articles vs. sentences around keywords for the category of *relevant* articles (in %. mean and standard deviation for five repeated runs).

The first finding is that with 88.4 % F1-score the out-of-sample set *test-time* scores significantly worse than the *test-within* set (93.3 %), but far less badly than the *test-loc* set. This is likely the case due to the greater similarity of linguistic features across time than across different cities. For the city of Leipzig, for instance, the protests in 2016 were largely dominated by anti-Muslim groups and counter-protesters, the patterns of which were highly repetitive. These protests began already throughout 2015. This might be the reason why our classifier is less negatively affected by the temporal out-of-sample dataset.

The second finding is that with 71.6 % F1-score the out-of-sample set *test-loc* performs considerably worse than the other two scenarios. Compared to the *test-within* set, performance drops by more than 20 pp. This drop can be attributed rather to a decrease in recall than in precision, which is actually not in favor of our application. We would prefer to rather remove false positives in the second step of our detailed manual annotation than miss out on relevant articles that are not identified as such due to lowered recall.

A third and pleasant finding is that we can indeed improve the classifier performance by focusing it around contexts of keyword matches. For all three test sets, we achieve higher performances by retaining only the sentences containing keywords and their direct neighbors (kwic+1). For the *test-time* set, we see a further but insignificant improvement by keeping an additional +2 sentences as broader contexts. For the *test-loc* out-of-sample scenario, the kwic+1 preprocessing improves the binary F1-score by around 4 pp. We, thus, consider our first hypothesis **H1** to be confirmed.

5.2. Masking Named Entities

We suspect that some share of over-fitting on our training data that leads to drastically decreased out-of-sample prediction can be attributed to spurious correlations. This may involve correlations of features that do not actually describe the protest events themselves, but that occur disproportionately often together with them. We hypothesize, that named entities are likely candidates for such features as local protest actors, organizations and places give hints for protest event detection that a machine classifier finds useful, but that do not generalize well across time and place.

To test this hypothesis, we apply a second variation of preprocessing to our so far best working approach: the fine-tuning of the ELECTRA model with articles reduced to the sentences around keyword matches (kwic+1). On this trimmed input text, we run named entity (NE) recognition using the `de-ner-large` model by (Akbik et al., 2019) that classifies tokens of a sentence into either a none-entity type or one of four entity types with the BIO-tagging schema. Then, token sequences tagged with a certain NE-type were replaced with a corresponding generic type token, i.e. 'Person' for `PER`-, 'Organisation' for `ORG`-, 'Ort' for `LOC`- and 'Name' for `MISC`-entities.

Again, we fine-tune the best performing transformer model *gelectra-large* analog to the previous experiment but with the replaced entities in the input texts: no masking of entities as a baseline, then masking of individual entity types and, last, masking of all types at once. Table 5 displays the results for all preprocessing variants on all three test datasets.

For *test-within* and *test-time*, we observe that compared to the baseline masking of individual or all entities has no significant impact on the overall performance. Only the masking of LOC-entities for the *within* set and ORG-entities for the *time* set seem to slightly improve precision but, at the same time, harm recall. For the *test-loc* set, we observe significantly improved precision when masking locations (+7.5 pp) but an even steeper drop in recall (-20 pp). For masking of organizations, the two metrics change in the opposite way. These results suggest that named entities have an ambiguous impact on protest event detection. In total, the insignificant positive effects on *test-time* (+0.3 pp F1-score for masking MISC entities), and the significant negative effects on *test-loc* (up to -10 pp F1-score for masking LOC entities) make clear that a simple replacement of named entities with generic terms seems to remove too much useful information for a machine classifier. We thus reject our second hypothesis **H2**. However, we notice positively that by masking all named entities, we can significantly increase the recall of relevant articles in the out-of-location sample (+16 pp) at the cost of precision (-17 pp).

Model	Precision	Recall	F1-score
		test-within	
no masking	91.7±1.4	**95.0±1.3**	93.3±0.3
mask PER	92.2±0.8	94.7±1.3	**93.4±0.4**
mask LOC	**92.8±0.8**	93.4±1.5	93.1±0.4
mask ORG	91.8±1.3	94.0±1.2	92.9±0.7
mask MISC	91.9±0.9	**95.0±0.9**	**93.4±0.5**
mask All	91.3±1.0	94.5±0.8	92.8±0.6
		test-time	
no masking	90.4±1.0	86.9±2.0	88.6±0.7
mask PER	89.7±0.9	88.0±3.1	88.8±1.2
mask LOC	90.6±1.0	85.3±4.2	87.8±1.9
mask ORG	**91.2±0.9**	85.5±2.5	88.3±0.9
mask MISC	89.7±0.8	**88.2±0.9**	**88.9±0.3**
mask All	89.8±1.3	87.4±0.5	88.6±0.5
		test-loc	
no masking	76.4±6.9	72.2±10.5	**73.5±2.8**
mask PER	73.0±7.5	74.0±10.2	72.7±2.2
mask LOC	**83.9±4.3**	52.2±11.1	63.5±7.9
mask ORG	57.9±6.4	**88.2±2.9**	69.7±3.9
mask MISC	80.2±12.7	62.9±21.2	67.1±8.2
mask All	59.4±3.6	88.0±2.0	70.9±1.9

Table 5: Classification performance with full-text articles vs. masking of named entities for the category of *relevant* articles (in %, mean and standard deviation for five repeated runs).

6. Conclusion

For the use in political science contexts, we introduced and evaluated the first German-language dataset for protest event detection in local news. It contains manually annotated data from four German cities and eleven consecutive years. The best transformer-based text classifier that we identified achieves a satisfactory in-sample binary F1-score of 93.3 % for automatic relevancy detection of news articles to support political scientists in their work on protest event analysis. For out-of-sample data from another city, however, this performance drops significantly due to drastically lowered precision. This effect can be mitigated by altering the preprocessing of articles for the classification process. We find that focusing the classification on sentences around protest-related keywords improves the out-of-sample F1-score for the *relevant* class up to 4 pp. A second method to improve the out-of-sample prediction performance, masking of named entities, did not prove to be successful.

Article relevancy classification is the first step in the pipeline of protest event detection within our project. We estimate that the approach presented in this paper can reduce the costs of human annotators for this step even in the hardest scenario of classifying articles from different places and different time periods by over 70 percent. This first step is followed by a step of detailed manual annotation and automatic classification of protest variables such as forms, claims, actors, and numbers of participants. For the targeted automation of this entire process to trace protest patterns across multiple cities and long time frames, the finding that masking of named entities can significantly improve the recall of relevant articles (at the cost of lowered precision) is actually still interesting for our future work. This way, we can filter out false positives in later steps of the pipeline while missing only few important articles at the initial step.

Acknowledgments

This work was funded by the German Federal Ministry of Education and Research (BMBF) as part of the project "Protests and Social Cohesion: Comparing Local Conflict Dynamics" within the *Research Institute for Social Cohesion* (RISC).

Bibliographical References

Beieler, J., Brandt, P. T., Halterman, A., Schrodt, P. A., and Simpson, E. M. (2016). Generating political event data in near real time. In R. Michael Alvarez, editor, *Computational Social Science*, pages 98–120. Cambridge University Press.

Buckland, M. and Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1):12–19.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia. OpenReview.net.

Croicu, M. and Weidmann, N. B. (2015). Improving the selection of news reports for event coding using ensemble classification. *Research & Politics*, 2(4):2053168015615596.

Dayanik, E. and Padó, S. (2020). Masking actor information leads to fairer political claims detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dollbaum, J. M. (2021). Protest event analysis under conditions of limited press freedom: Comparing data sources. *Media and Communication*, 9(4):104–115.

Gladun, A. (2020). Protesting that is fit to be published: issue attention cycle and nationalist bias in coverage of protests in Ukraine after Maidan. *Post-Soviet Affairs*, 36(3):246–267.

Hanna, A. (2017). MPEDS: Automating the generation of protest event data. Technical report, SocArXiv.

Hürriyetoğlu, A., Yörük, E., Yüret, D., Yoltar, Ç., Gürel, B., Duruşan, F., and Mutlu, O. (2019). A task set proposal for automatic protest information collection across multiple countries. In Leif Azzopardi, et al., editors, *Advances in Information Retrieval*, volume 11438 of *Lecture notes in computer science*, pages 316–323. Springer International Publishing, Cham.

Hürriyetoğlu, A., Zavarella, V., Tanev, H., Yörük, E., Safaya, A., and Mutlu, O. (2020). Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 1–6, Marseille, France. European Language Resources Association.

Hutter, S. (2014). Protest event analysis and its offspring. In Donatella Della Porta, editor, *Methodological Practices in Social Movement Research*, pages 335–367. Oxford University Press.

Hürriyetoğlu, A., Mutlu, O., Yörük, E., Liza, F. F., Kumar, R., and Ratan, S. (2021). Multilingual Protest News Detection - Shared Task 1, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 79–91, Online, August. Association for Computational Linguistics.

Koopmans, R. and Rucht, D. (2002). Protest event analysis. In *Methods of Social Movement Research*, page 231–259. University of Minnesota Press.

Lorenzini, J., Makarov, P., and Wüest, B. (2020). Design and methods of semi-automated protest event analysis. In Hanspeter Kriesi, et al., editors, *Contention in Times of Crisis: Recession and Political Protest in Thirty European Countries*, pages 29–48. Cambridge University Press.

Nam, T. (2006). What you use matters: Coding protest data. *Political Science & Politics*, 39(2):281–287.

Nardulli, P. F., Althaus, S. L., and Hayes, M. (2015). A progressive supervised-learning approach to generating rich civil strife data. *Sociological Methodology*, 45(1):148–183.

Raleigh, C., Linke, A., Hegre, H., and Karlsen, J. (2010). Introducing acled: An armed conflict location and event dataset: Special data feature. *Journal of Peace Research*, 47(5):651–660.

Schrodt, P. A. and Van Brackle, D. (2013). Automated coding of political event data. In V.S. Subrahmanian, editor, *Handbook of Computational Approaches to Counterterrorism*, pages 23–49. Springer, New York, NY.

Wang, W., Kennedy, R., Lazer, D., and Ramakrishnan, N. (2016). Growing pains for global monitoring of societal events. *Science*, 353(6307):1502–1503.

Weidmann, N. B. and Rød, E. G. (2019). *The Internet and Political Protest in Autocracies*. Oxford University Press.

Wiedemann, G. and Fedtke, C. (2021). From frequency counts to contextualized word embeddings: The Saussurean turn in automatic content analysis. In Uwe Engel, et al., editors, *Handbook of Computational Social Science, Volume 2*, pages 366–385. Routledge, London.

Wiedemann, G. (2017). Vertrauen und Protest: Eine exemplarische Analyse des Demonstrationsgeschehens in der BRD mithilfe von Text Mining in diachronen Zeitungskorpora. In Michael Haller, editor, *Öffentliches Vertrauen in der Mediengesellschaft*, pages 172–200. Herbert von Halem Verlag, Köln.

Wüest, B. and Lorenzini, J. (2020). External validation of protest event analysis. In Hanspeter Kriesi, et al., editors, *Contention in Times of Crisis: Recession and Political Protest in Thirty European Countries*, pages 49–78. Cambridge University Press.

Zhang, H. and Pan, J. (2019). CASM: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, 49(1):1–57.

Language Resource References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Chan, B., Schweter, S., and Möller, T. (2020). German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., and Gurevych, I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of*

the 27th International Conference on Computational Linguistics: System Demonstrations, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Wiedemann, G., Dollbaum, J. M., Haunss, S., Daphi, P., and Meier, L. D. (2022). German Local Protest News (GLPN) Dataset. Distributed via Zenodo, DOI: 10.5281/zenodo.6490537.